15400 Calhoun Drive, Suite 400
Rockville, Maryland, 20855
(301) 294-5200
http://www.i-a-i.com

Intelligent Automation Incorporated

Information Tailoring Enhancements for Large-Scale Social Data

Progress Report No. 3

Reporting Period: March 16, 2016 – June 15, 2016

Contract No. N00014-15-P-5138

Sponsored by
ONR, Arlington VA
COTR/TPOC: Dr. Rebecca Goolsby

Prepared by
Onur Savas, Ph.D.

DISTRIBUTION A

Approved for public release; distribution is unlimited.

Progress Report No. 3

Information Tailoring Enhancements for Large-Scale Social Data

Submitted in accordance with requirements of
Contract #N00014-15-P-5138

Performance period: March 16, 2016 to June 15, 2016
(PI: Dr. Onur Savas, 301.294.4241, osavas@i-a-i.com)

1 Work Performed within This Reporting Period

In this reporting period, we performed the following tasks.

- **Enhanced Named Entity Recognition Capabilities:** We have enhanced the Named Entity Recognition (NER) capabilities of Scraawl by incorporating (i) part-of-speech tagging using GATE, (ii) enhancing the gazetteers with multilingual entities, and (iii) adding multi-lingual name matching capabilities. These capabilities will be made available in Scraawl advanced analytics no later than the end of August, 2016.

- **Delivered Scraawl Version 1.15.**

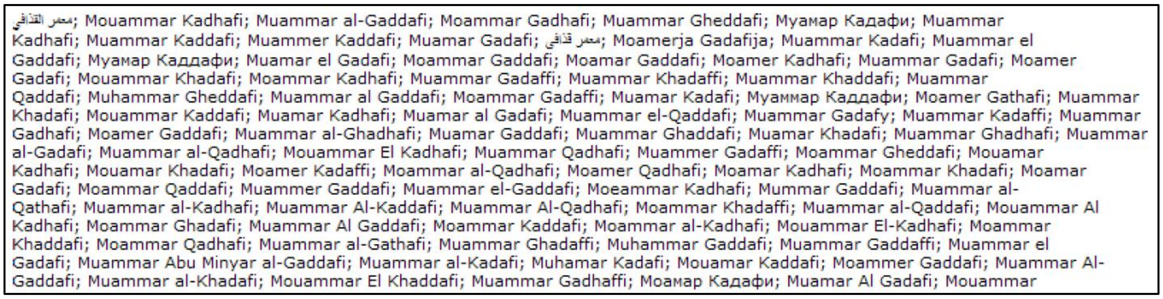
1.1 Enhanced Named Entity Recognition (NER) Capabilities

As of version 1.15, Scraawl has NER capabilities of resolving a large set of names, organizations, and places in English. It also expands organization abbreviations, e.g., US to United States. During this reporting period, we have made the following improvements to the Scraawl NER module, which will be made available in Scraawl advanced analytics no later than the end of August, 2016.

Incorporated GATE Part-of-Speech (POS) tagging: We have started using General Architecture for Text Engineering (GATE) software’s [1] English POS tagger as part of Scraawl NER module. GATE is an open source software to do many common task related to Natural Language Processing. Its POS tagger [2] is a modified version of the

Brill tagger, which produces a part-of-speech tag as an annotation on each word or symbol. The current NER development software uses classifies a word as an entity if and only if the word is one of the gazetteers and its POS tag is Noun.

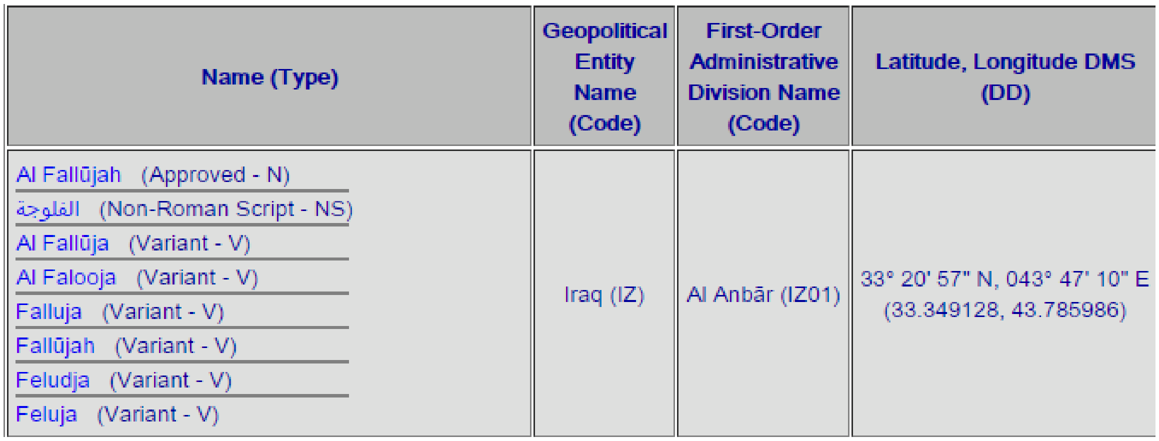
Enhanced the gazetteers and incorporated multi-lingual name matching capabilities: We have enhanced the gazetteers by including open source JRC-Names dictionary [3], and NGA Geographical Names Database [4]. JRC-Names contains the most important names of the EMM name database, i.e., those names that were found frequently or that were verified manually or found on Wikipedia. In particular, the Europe Media Monitor (EMM) family of applications gather a current average of 100,000 news articles per day in up to 50 languages from the internet, classify them into hundreds of categories, cluster related news, link news clusters over time and across languages, and – for twenty languages – perform entity recognition, classification and disambiguation for the entity types person, organization and location. EMM also gathers information about entities from all news articles and displays it on over one million entity pages [5][6], and the information is made available in JRC-Names. The compiled dictionary. Which has 1.18+ million person and 6700+ organizations, has variants of names and organizations as well. A representative example of name variant spellings for Libyan leader Muammar Gaddafi, as found in multilingual media reports [5] is depicted in Figure 1. Including name variants in the dictionary allows us to perform basic name matching and entity resolution.



معمر القذافي; Mouammar Kadhafi; Muammar al-Gaddafi; Moammar Gadhafi; Muammar Gheddafi; Муамар Кадафи; Muammar
Kadhafi; Muammar Kaddafi; Muammer Kaddafi; Muamar Gadafi; معمر قذافي; Moamerja Gadafija; Muammar Kadafi; Muammar el
Gaddafi; Муамар Каддафи; Muamar el Gadafi; Moammar Gaddafi; Moamar Gaddafi; Moamer Kadhafi; Muammar Gadafi; Moamer
Gadafi; Mouammar Khadafi; Moammar Kadhafi; Muammar Gadaffi; Muammar Khadaffi; Muammar Khaddafi; Muammar
Qaddafi; Muhammar Gheddafi; Muammar al Gaddafi; Moammar Gadaffi; Muamar Kadafi; Муаммар Каддафи; Moamer Gathafi; Muammar
Khadafi; Mouammar Kaddafi; Muamar Kadhafi; Muamar al Gadafi; Muammar el-Qaddafi; Muammar Gadafy; Muammar Kadaffi; Muammar
Gadhafi; Moamer Gaddafi; Muammar al-Ghadhafi; Muamar Gaddafi; Muammar Ghaddafi; Muamar Khadafi; Muammar Ghadhafi; Muammar
al-Gadafi; Muammar al-Qadhafi; Mouammar El Kadhafi; Muammar Qadhafi; Muammer Gadaffi; Moammar Gheddafi; Mouamar
Kadhafi; Mouamar Khadafi; Moamer Kadaffi; Moammar al-Qadhafi; Moamer Qadhafi; Moamar Kadhafi; Moammar Khadafi; Moamar
Gadafi; Moammar Qaddafi; Muammer Gaddafi; Muammar el-Gaddafi; Moeammar Kadhafi; Mummar Gaddafi; Muammar al-
Qathafi; Muammar al-Kadhafi; Muammar Al-Kaddafi; Muammar Al-Qadhafi; Moammar Khadaffi; Muammar al-Qaddafi; Mouammar Al
Kadhafi; Moammar Ghadafi; Muammar Al Gaddafi; Moammar Kaddafi; Moammar al-Kadhafi; Mouammar El-Kadhafi; Moammar
Khaddafi; Moammar Qadhafi; Muammar al-Gathafi; Muammar Ghadaffi; Muhammar Gaddafi; Muammar Gaddaffi; Muammar el
Gadafi; Muammar Abu Minyar al-Gaddafi; Muammar al-Kadafi; Muhamar Kadafi; Mouamar Kaddafi; Moammer Gaddafi; Muammar Al-
Gaddafi; Muammar al-Khadafi; Mouammar El Khaddafi; Muammar Gadhaffi; Моамар Кадафи; Muamar Al Gadafi; Mouammar

Figure 1: Variants of Muammar Gaddafi as represented in JRC-Names [5].

NGA Geographical Names Database [4] is a multi-lingual compilation (with dialects and variety of common spellings) of National Geospatial-Intelligence Agency's (NGA) and the U.S. Board on Geographic Names' (BGN) database of foreign geographic feature names. The data is in a geographic coordinate system based on the WGS84 datum and ellipsoid. Geographic coordinates are approximate and are intended for finding purposes. The online database is updated weekly. We have incorporated this gazetteer as part of the Scraawl NER and also use it to perform location matching in multiple languages. A representative example is shown in Figure 2.



Name (Type)	Geopolitical Entity Name (Code)	First-Order Administrative Division Name (Code)	Latitude, Longitude DMS (DD)
Al Fallūjah (Approved - N) الفلوجة (Non-Roman Script - NS) Al Fallūja (Variant - V) Al Falooja (Variant - V) Falluja (Variant - V) Fallūjah (Variant - V) Feludja (Variant - V) Feluja (Variant - V)	Iraq (IZ)	Al Anbār (IZ01)	33° 20' 57" N, 043° 47' 10" E (33.349128, 43.785986)

Figure 2: Representative NGA Location Entry for Fallujah.

2 Current Problems

None.

3 Work to be Performed in the Next Reporting Period

In the next report period, we will focus on the following tasks:

- We will enhance geo-reference analytics.
- We will deliver Scraawl 1.16.

4 Financial Status

Financially, we are in good shape.

References

[1] GATE, https://gate.ac.uk.

[2] M. Hepple. Independence and commitment: Assumptions for rapid training and execution of rule-based POS taggers. In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000), Hong Kong, October 2000.

[3] JRC-Names, https://ec.europa.eu/jrc/en/language-technologies/jrc-names.

[4] NGA Geographic Names Database, https://www.nga.mil/ProductsServices/GeographicNames/Pages/default.aspx.

[5] Steinberger Ralf, Bruno Pouliquen, Mijail Kabadjov, Jenya Belyaeva & Erik van der Goot (2011). JRC-Names: A freely available, highly multilingual named entity.Proceedings of the 8th International Conference Recent Advances in Natural Language Processing (RANLP). Hissar, Bulgaria, 12-14 September 2011.

[6] Steinberger Ralf, Bruno Pouliquen & Erik van der Goot (2009). An introduction to the Europe Media Monitor family of applications. Proceedings of the SIGIR'2009 Workshop 'Information Access in a Multilingual World'. Boston, USA.

.